DOCUMENT RESUME

ED 046 600 24 RC 005 014

AUTHOR Havighurst, Robert J.
TITLE The Reliability of Pating Scales Used in Analyzing Interviews with Parents, Students, Teachers, and Community Leaders. The National Study of American Indian Education, Series IV, No. 9, Final Report.
INSTITUTION Chicago Univ., Ill.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO BR-8-0147
PUB DATE Dec 70
CONTRACT OEC-0-8-080147-2805
NOTE 11p.

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS *American Indians, Attitudes, Community Surveys, *Comparative Analysis, Correlation, Education, Interviews, *National Surveys, Parents, *Rating Scales, *Reliability, Schools, Students, Teachers, Validity

ABSTRACT

As part of the Final Report of the National Study of American Indian Education, this paper reports on the reliability of rating scales used in analyzing the interviews conducted during the study. Approximately 4,000 interviews, which covered "adequate samples of people in the schools and communities" that were studied, were deemed valid as a source of accurate data. The rating scales devised by various field centers to analyze these interviews were used (1) to evaluate a particular school or school system of a particular community, (2) to compare schools and communities singly and in various combinations, and (3) to compare perceptions and attitudes of parents with students, parents with teachers, teachers with students, etc. The document provides an explanation of the components that make up the rating scales and the results. Reliability of the instruments used and the procedures used to study reliability are also discussed. It is concluded that reliability of the ratings from the various field centers was high enough to allow for comparisons between various schools or communities and between various types of respondents to the interview. (EL)

The National Study of American Indian Education

USOE Project No. OEC-0-8-08147-2805 FINAL REPORT

Series IV. No. 9 December, 1970

THE RELIABILITY OF RATINC SCALES USED IN ANALYZING INTERVIEWS WITH PARENTS, STUDENTS, TEACHERS, AND COMMUNITY LEADERS

Robert J. Havighurst

NATIONAL STUDY OF AMERICAN INDIAN EDUCATION

The attached paper is one of a number which make up the Final Report of the National Study of American Indian Education.

This Study was conducted in 1968-69-70 with the aid of a grant from the United States Office of Education, OEC-0-8-080147-2805.

The Final Report consists of five Series of Papers:

I. Community Backgrounds of Education in the Communities Which Have Been Studied.

II. The Education of Indians in Urban Centers.

III. Assorted Papers on Indian Education--mainly technical papers of a research nature.

IV. The Education of American Indians--Substantive Papers.

V. A Survey of the Education of American Indians.

The Final Report Series will be available from the ERIC Document Reproduction Service after they have been announced in Research in Education. They will become available commencing in August, 1970, and the Series will be completed by the end of 1970.

THE RELIABILITY OF RATING SCALES USED IN ANALYZING INTERVIEWS WITH PARENTS, STUDENTS, TEACHERS, AND COMMUNITY LEADERS

In the National Study we have relied very heavily on interviews as a principal source of information. We have held almost four thousand interviews with fairly adequate samples of people in the schools and communities that we studied. We have evidence of the validity of these interviews as a source of accurate data. The interview procedure and its validity have been described and discussed in papers 7 and 8 of this series.

We have analyzed the interviews quantitatively with the aid of 63 rating scales. Some of these scales have proved much more useful than others in our subsequent interpretations and conclusions. Approximately 25 of them have been used most heavily in the papers that make up this series (Numbers 5, 10, 11, and 12).

The rating scale data have been used for the following purposes:

(1) To evaluate a particular school or the school system of a particular community.
(2) To compare schools and communities singly and in various combinations.
(3) To compare the perceptions and attitudes of parents with students, parents with teachers, teachers with students, etc.

The second and third purposes, which involve comparison of data from several groups or sets of interviews, require data which are reliable as well as valid. In order to use ratings for comparison purposes, it is necessary to establish the consistency or the reliability of these ratings. That is, it is necessary to show that the judges or raters of a given center give the same ratings on a particular interview as the judges or raters of another center. It is, of course, also necessary to establish consistency or reliability of the ratings made by two or more judges within a center.

There are a number of sources of error or disagreements by judges or raters. Some of these are:

Halo-effect--rating the interviews favorably or unfavorably because the judge sees the situation himself favorably or unfavorably.

Leniency--rating the interviews favorably through giving the interview the "benefit of the doubt" because the judge knows the situation being rated and supposes that the respondent would see things that way if he knew more about the situation.

Clustering ratings near the center of the scale--tendency of some judges to be very conservative about giving very high or very low ratings.

Logical error--giving the same rating for traits or dimensions that seem to the judge to be logically related, though this fact has not been established.

Some of these errors can be reduced by skillful construction of rating scales. Others can be reduced by training raters or judges to be more careful and more objective in their ratings.

In the scales developed for this study, some time and effort were spent to reduce the probability of errors:

1. Members of several center staffs worked together in constructing the rating scales; discussing the meanings of the various dimensions and the various scale points on a given dimension.

2. Sample interviews were studied in constructing rating scales.

3. Sample interviews were rated with preliminary rating scales. The judges or raters compared these ratings, and then revised the rating scales to clear up ambiguities where they had disagreed in their ratings.

In the end, each rating scale was examined and revised at least once by two or more people using the procedures noted above.

Reliability of the Revised Scales

When the rating scales had finally been sent out for use by the various centers, the time had come to study the reliability or consistency of the ratings from any one center, and also to study the cross-center reliability.

Basically, this process required a statistical comparison of the ratings of 2 or more judges on a number of interviews which were rated by these judges.

The procedure we adopted was a 2-stage process:
(1) Testing reliability of the scales and raters within a field center;
(2) Testing reliability of the scales and raters between field centers.

It was decided that 21 interviews of a given type were sufficient for this test. When we came to the problem of cross-center reliability, this would permit us to use 3 interviews from each of the seven field centers.

Intra-Center Reliability. The procedure worked out for the study of reliability within a center was as follows:

1. Each center assign the rating of a given type of interview to 2 or more staff members.
2. These staff members rate several interviews and improve their ratings by discussing disagreements and working out "ground rules."
3. A sample of 10 or more interviews should be chosen. These should be representative of the interviewers and respondents in the center.
4. Each rater from the center should rate all of these interviews--independently.
5. The different raters should compare their respective ratings. Where there is disagreement of 2 or more steps on a 7-point scale, they should discuss their differences and try to work out a way of rating which eliminates this kind of discrepancy.
6. When they eventually reach a satisfactory degree of agreement, they should go ahead and rate all of their interviews, sending a report of their ratings to Chicago, where comparative studies are being made.

. . . .

There is obviously an essential step missing in this procedure, if comparisons of data are to be made from several field centers. It must be established that the ratings made in one Center are consistent with and substantially similar in basic interpretation to the ratings from other Centers.

That is, it must be established that the judges in one Center are operating with the same interpretations and "general rules" as those of another Center. Then, if there are differences between the ratings of interview from Center A and Center B, these differences reflect real differences in content of the interviews, rather than differences in the application by the raters or judges of the rating scales. Therefore, it is necessary to establish the inter-center reliability of the ratings, as well as the intra-center reliability.

Inter-Center Reliability

The procedure for measuring the inter-center reliability of ratings was for each center to rate the same set of interviews that was rated by all the other centers. Then the reliability coefficients were computed for the group of ratings, counting each Center as a single rater or judge. The set of interviews was arbitrarily defined to consist of 3 interviews for each of the seven Centers. These should be as representative as possible: representative of types of parents, ages of students, sex of teachers, types of interviewer, etc.

There were four sets of 21 interviews; one set for each type of respondent. Each set of 21 interviews was duplicated and sent to each Center to be rated there by two or more judges. Their ratings were reported to the Chicago office, which compared the ratings and computed the reliability indices.

The Reliability Computations. Assume we have 21 interviews which have been rated on Dimensions or Variables A, B, C, D, ... For each variable there is a rating scale of 5 to 7 steps. The ratings are not spread evenly over these steps, but tend to cluster at middle values, for most of the variables.

The 21 interviews have been rated by teams at 6 Centers. (There were seven Centers, but one did not participate in the reliability test. Its staff made their ratings later, but their ratings compared acceptably with those of the Chicago Center, on a test group of interviews.) Thus, we have 21 interviews, each rated by 6 "raters" or "judges" (counting the average rating coming from a given center as being made by one judge).

The next step is to apply appropriate statistical tests of reliability. Three such tests are used. With 21 interviews, we will generally have a number of interviews with the same scale value--e.g., six interviews rated at level 5, five at level 4, five at level 3, etc. These are really ties, and several statistical reliability measures cannot be used in this simple form, which assumes no ties.

Thus, Kendall's tau coefficient can only be used with ties with a special formula. Spearman's rank order coefficient should not be used with ties, but it is often used with a small number of ties, since the error so introduced is not great.

The statistical entities we might use are:
Spearman's rank order correlation coefficient (in case of only a small number of ties)
Kendall's tau coefficient, corrected for ties
Goodman and Kruskal's gamma coefficient
Kendall's coefficient of concordance, W, corrected for ties
The Pearson product-moment correlation coefficient
The intraclass correlation coefficient

One of the more thorough discussions of the problem of the reliability of ratings is given by Guilford, (1) and his suggestions are followed in this paper.

Counting agreements and one- and two-step disagreements among judges or raters.

A simple method of measuring reliability and one which keeps the actual data in sight for us to look at is to count the numbers of agreements among 2 or more raters, the numbers of one-step disagreements, the numbers of 2-step disagreements, and so on. For example, with a 6-point rating scale and each judge assigning scale values at random, there would be 1 in 6 or 16 percent agreement between a pair of judges, 16 percent 1-step disagreement, etc. Thus a higher proportion of agreement than 16 percent might be taken as evidence of reliability. But this would seldom be accurate, because on a rating scale very often the extreme scale points are very seldom used by judges, and therefore the actual spread of ratings is narrowed down, and the chance of agreement by random ratings is increased. For instance, if nearly all of the ratings made by judges were placed at points 4 and 5 of a 6-point scale, the judges would agree nearly half of the time if they used a random method of choosing between ratings 4 and 5.

However, most scales are designed so as to get a fairly wide distribution of ratings. When we get as much as 80 percent agreement or one-step disagreement between pairs of judges we can be pretty sure that the agreement is much greater than we could expect by chance.

When we have six judges or raters, as we have in studying the inter-Center reliability, complete agreement of all six judges or a maximum of one-step or two-step disagreement among the judges in rating a given dimension is extremely improbable on the basis of chance alone. Almost all of our ratings are consistent enough among the various centers to be reliable at far beyond the chance level. We will report these computations on two of the rating scales as an example.

Computing Coefficients of Reliability. A more sophisticated procedure to study reliability or consistency of ratings is to compute one or another type of coefficient of reliability. The ones we have used range from 0 (no reliability) to 1 (complete consistency or reliability). They all can be interpreted crudely as one would interpret a product-moment correlation coefficient. That is, coefficients of less than.50 indicate unsatisfactory reliability, and coefficients of .70 or more indicate a very satisfactory reliability.

With ratings from 6 judges, we computed the following coefficients:

1. Product-moment (Pearson) correlation coefficient between pairs of judges.
2. Kendall's Coefficient of Concordance (W), corrected for ties in the ratings.
3. Intraclass correlation coefficient for one pair and a group of several judges--a method based on analysis of variance.
4. Kendall's tau coefficient, corrected for ties in the ratings.
5. Goodman and Kruskal's gamma coefficient. (This can be done easily in the course of computing Kendall's tau.)

Eventually we settled on numbers 1, 2, and 3 of the list named above. These told us all we needed to know. Furthermore, we computed these for only some, not all, of the dimensions being rated. There was no advantage to doing the extra computation on the ratings for a new dimension when inspection of the ratings showed a vcry similar pattern of agreement and disagreement to another dimension which had already been thoroughly studied. The counting of agreements and one- and two-step disagreements was enough to assure us that we would get reliable correlation coefficients for all dimensions that presented similar patterns of agreement between judges.

Table 1 shows the criteria we adopted for concluding that the ratings on a given Dimension or Scale have high, medium, or low reliability. As a matter of fact, even the scales with medium reliability have coefficients that are far above the chance level.

Table 2 shows the level of reliability of the ratings on the scales which have been used in papers 5, 10, 11, and 12 of this series. It will be noted that the only scales with a "low" reliability level were some from the interviews with community leaders. The interviews with these people were not carried through as thoroughly and carefully as were the other types of interviews. We have been rather cautious in basing our conclusions or interpretations on these ratings.

Comments on Specific Reliability Measures

The product-moment correlation coefficient. This is a measure of agreement between any pair of raters or judges. It is a well-known statistic, and we think it useful for that reason. A disadvantage is that our rating scales are so short (6 to 7 points) that we do not get much differentiation of scores. Also, the number of interviews is relatively small (21 in our model procedure), thus allowing a substantial probable error. Still, this procedure gets away from the problem of tied scores.

The Spearman rank order coefficient (rho). This, also, is well-known, and gives a measure of agreement between any two judges. However, it loses accuracy when tied scores are present, as is the case with many of our ratings. We therefore have not used it very often.

Kendall's Coefficient of Concordance (W)[3] [2] This is a useful measure of agreement among 2 or more raters, and it has a correction procedure for tied ranks. We believe this is a better measure than the two preceding ones.

Intraclass Correlation Coefficient.[1] This is a procedure based on analysis of variance, which gives a reliability coefficient for an average judge in a group of 2 or more judges. It also gives a reliability coefficient for the average of 2 or more judges. There is no problem of correction for ties. This is probably the most useful of the coefficients we have calculated.

Example of the Reliability Report on a Specific Dimension

The following paragraphs contain reports on the reliability of ratings by 6 Field Centers on two rating scales. In the first example, the reliability is high and in the second example the reliability is described as medium, although it is far beyond the level of a chance assignment of ratings.

It will be noted that the highest coefficients come from the Coefficient of Concordance (W) and from the intraclass correlation R_{66} for six judges. This is to be expected, for the coefficients give the reliability of the combined ratings of the six judges. For any given pair of judges, (or of Centers) the reliability would be lower.

Student Interview Dimension O. Respondent's Opinion of his Teacher(s)

Complete agreement among all 6 Centers in 7 or 33 percent of the cases.
One-step disagreement among the raters; spread of ratings is only 2 steps in 10 or 47 percent of the cases.
Two-step disagreement among the raters; spread of ratings is 3 steps in 4 or 20 percent of the cases.
Product-moment correlation coefficients between pairs of Centers: .49, .87, and .50 for three pairs.
Kendall's Coefficient of Concordance: (W) corrected for ties, .71.
Intraclass correlation: average of two raters, .69; average of 6 raters, .93.
Reliability is high

Community Leader Interview B. Respondent's Overall Evaluation of the School Program

Complete agreement among all 6 Centers in 0 percent of cases.
One-step disagreement among the raters: spread of ratings is only 2 steps in 7 or 33 percent of the cases.
Two-step disagreement among the raters: spread of ratings is three steps in 11 or 53 percent of the cases.
Three-step disagreement among the raters: spread of ratings is four steps in 3 or 14 percent of the cases.
Product-moment Correlations between pairs of Centers: .49, .64, .06 for three pairs.
Kendall's Coefficient of Concordance (W) corrected for ties, .70.
Intraclass correlation: Average of two raters, .43; Average of 6 raters, .86.
Reliability is medium

Corrections for Differences in Level of Ratings

The preceding pages have documented the relatively high reliability of the data for the rating scales. However, they are incomplete in that they do not deal directly with another possible source of discrepancy between raters--one of level of rating. If one rater always rates interviews one point higher than another, they will show a perfect correlation, although their ratings are systematically different. Therefore a comparison of ratings from a "high rater" with those from a "low rater" would lead one to expect real differences between two schools, whereas the real difference lay in the rating procedures. This source of error can be eliminated by making systematic corrections of the ratings by "high" and "low" raters.

The corrections are made by comparing the average ratings for the 21 sample interviews on a given scale or dimension made by the judges of a given center with the average on that scale from the other five centers. If there is a difference between these two figures, this difference becomes the correction to be applied to the scale that should be corrected. We find, for example, corrections on Dimension O of the Student Interview (where the ratings were highly reliable) ranging from plus .24 for one Center to minus .33 for another Center. Three of the Centers did not require corrections. Corrections have been applied to the interview ratings that are reported in papers 5, 10, 11, and 12.

Conclusion

The consistency of rating and the reliability of the ratings from the various field centers appear to be so high as to permit useful comparisons between various schools or communities, and between various types of respondents to the interview. We have limited our comparisons to dimensions of the interviews which have been shown to be reliable to a satisfactory degree.

REFERENCES

1. Guilford, J. P. Psychometric Methods. esp. pp. 395-397. New York: McGraw Hill, 1954.

2. Edwards, Allen. Statistical Methods for the Behavioral Sciences. pp. 402ff. and 430-433. New York: Rinehart. 1954.

3. Kendall, Maurice G. Rank Correlation Methods. pp. 94-96. London: Griffin. 1955.

Table 1

CRITERIA FOR DEFINITION OF RELIABILITY LEVELS OF RATING SCALES USED FOR INTER-SCHOOL COMPARISONS

Reliability Coefficients	High	Medium	Low
Kendall's Coefficient of Concordance(W) for 6 Raters and 21 cases	.70 plus	.6 to .7	Below .6
Intra-Class Correlation for 1 Pair of Raters and 21 cases	.60 plus	.5 to .6	Below .5
6 Raters and 21 cases	.92 plus	.86 to .91	Below .85
Average Product-Moment Correlation Coefficient between Pairs of Raters, for 21 cases	.5 plus	.3 to .5	Below .3
Percent of 21 Cases with Ratings from 6 Centers which show:			
Complete Agreement	10 plus	0 to 10	0
Maximum of 1-step disagreement between 2 or more Raters	40 plus	30 to 50	Below 25
Maximum of 2-step disagreement between 2 or more Raters	10 to 40	25 to 45	25 to 45
Maximum of 3-step disagreement between 2 or more Raters	0	0 to 20	20 plus

Table 2

INTER-CENTER RELIABILITY LEVEL FOR RATING DIMENSIONS USED FOR INTER-SCHOOL COMPARISONS

	Scales We Used	Inter-Center Reliability Level		
		High	Medium	Low
Parent				
I-A	Parent's Knowledge of School's Program and Policy	X		
II-B	Parent's Perception of How Well the School is Meeting the Needs of his Child	X		
II-C	Parent's Attitude Toward Formal Education		X	
IV-A	Parent's Involvement in School Affairs	X		
V-C	Attitude Toward Teaching Tribal History and and Culture in the School	X		
VI-A	Parent's Opinion of His Child's Teacher(s)	X		
VI-B	Parent's Opinion of the Curriculum in His Child's School	X		
VI-C	Parent's Opinion of the Performance of the School Administration	X		
Student				
J.	Attitude Toward School's Relationship to Tribal Culture		X	
K.	Respondent's Opinion of the School He is Now Attending	X		
L.	Student's Interest in the Academic Aspect of School		X	
O.	Respondent's Opinion of his Teacher(s)	X		
Teacher				
A.	Teacher's Experience and Knowledge of the Local Community	X		
B.	Teacher's Degree of Understanding of and Sympathy for the Problems of Local Indian People	X		
C.	Attitude Toward Assimilation versus Maintaining a Separate Indian Culture	X		
F.	Teacher's Attitude Toward Teaching Indian Children	X		
Community Leader				
B.	Respondent's Overall Evaluation of School Program		X	
D.	Respondent's Perception of the Effectiveness of the School in Assisting Students Toward Effective Participation in Modern Society			X
E.	Respondent's Attitude Toward Teaching Indian History and Culture in the School		X	
J.	Respondent's Attitude Toward Local Community Control over School		X	
I.	Respondent's Perception of Local Indian and Community Influence on the School Program			X